Qeios

Peer Review

Review of: "A Comparative Study of Large Language Models in Explaining Intrinsically Disordered Proteins"

Vinay Sammangi[1]

1. Georgia Institute of Technology, United States

The study evaluates the performance of four large language models (GPT-3.5, GPT-4, GPT-4 with Browsing, and Google Bard) in explaining intrinsically disordered proteins (IDPs). Using five evaluation metrics - accuracy, relevance, depth, clarity, and quality—the authors highlight GPT-4 as the leading performer. The study underscores the potential of AI tools like GPT-4 in advancing scientific education and bridging gaps in niche domains such as IDPs.

Strengths:

- Focuses on leveraging AI for a specialized, challenging domain like IDPs.
- Expert-blinded assessments coupled with statistical analysis enhance credibility.
- Highlights GPT-4's superior performance and identifies areas for improving AI in education.
- Well-organized tables and figures effectively summarize key findings.

Suggestions for Improvement:

- Explain why GPT-3.5 occasionally performed better than GPT-4 in 2 cases.
- Address variability in AI responses by analyzing if models produced consistent responses, given close performance scores.
- Clarify Figure 3 Metrics: specify how the "best model" frequency was calculated (average across five metrics?).
- Use more than 10 questions and 5 use cases for broader, better comparisons.
- Eliminate redundant text by consolidating the evaluation criteria explanation in either Table 1 or in the paragraph before that (not in both).

doi.org/10.32388/UB3LCT

- Correct Table 1 by including the missing "rating scale" column referenced in the text.

- Update Table 4 "shading for statistically significant differences" or revise the text.

Declarations

Potential competing interests: No potential competing interests to declare.